On the Impact of Garbage Collection on flash-based SSD endurance

Robin Verschoren and Benny Van Houdt

*Dept. Mathematics and Computer Science
University of Antwerp
Antwerp, Belgium

INFLOW '16

Universiteit Antwerpen

Outline

- SSD basics
- Prior work
- System description
  - GC algorithms
  - Workloads
  - GC mechanism
- Performance measures
- Model framework
- Findings
- Future work

Universiteit Antwerpen

Flash-based SSD

SSD Structure (plane level)

- Data is organized in N blocks
- Fixed number of b pages per block (e.g., $b = 32$)
- Unit of data exchange is a page
- Page has 3 possible states: erase, valid or invalid.

Operations

- Data can only be written on pages in erase state
- Erase operations can be performed on entire blocks only
- Each block tolerates W_{max} program-erase (PE) cycles before wearing out
- Out-of-place writes are supported (old data becomes invalid)

Universiteit Antwerpen

Flash-based SSD

Internal operation (internal log structure)

- New data is sequentially written to one or more special blocks called write frontiers (WFs)
- When a WF is full, a new WF is selected by the garbage collection (GC) algorithm

Write Amplification (WA)

- Valid pages in the victim block are temporarily copied to perform erase
- Assume j valid pages on a victim block with probability p_j, write amplification A equals

$$A = \frac{b}{b - \sum_{j=0}^{b} j p_j}$$

Write Amplification

Importance

- Affects IOPS and life span of the drive

Over-provisioning

- Physical storage capacity exceeds the user-visible (logical) capacity
- Measure is spare factor $S_f = 1 - \rho$:

$$\rho = \frac{\text{the user-visible capacity}}{\text{total storage capacity}}$$

$\Rightarrow$ fraction S_f of the pages is guaranteed to be in erase/invalid state

Prior work

Analytical models

- Mostly under uniform random writes and Rosenblum (hot/cold) workloads
- Exact (closed-form) results for WA as N tends to infinity
  - Random GC
  - FIFO/LRU GC (Menon, Robinson, Desnoyers)
  - Greedy GC (Bux, Iliadis, Desnoyers)
  - d-choices GC (Van Houdt, Li et al.)
  - Windowed GC (Hu et al., Iliadis)
  - etc.

Universiteit Antwerpen

Prior work

Main observations w.r.t. Write Amplification

- Greedy is optimal under uniform random writes, *d*-choices close to optimal (for *d* as small as 10)
- Increasing hotness worsens WA in case of single WF (as no hot/cold data separation takes place)
- Double WF (separates writes triggered by host and GC): WA decreases as hotness increases (as partial hot/cold data separation takes place)
- Greedy is no longer optimal with hot/cold data: there exists optimal *d* for *d*-choices

Universiteit Antwerpen

Class $\mathcal{C}$ of GC algorithms modeled

Definition

- Let $\vec{m}(t) = (m_0(t), \ldots, m_b(t))$, where $m_i(t)$ is the fraction of blocks containing i valid pages at time t
- A GC algorithm belongs to $\mathcal{C}$ if
  1. A block containing j valid pages is selected by the GC algorithm with probability $p_j(\vec{m}(t))$
  2. The probabilities $p_j(\vec{m}(t))$ are smooth in $\vec{m}(t)$ (can be slightly relaxed)
- It is possible to further extend this class when hot/cold data identification techniques are in place

Universiteit Antwerpen

Class $\mathcal{C}$ of GC algorithms modeled

Examples

1. Random GC algorithm: $p_j(\vec{m}) = m_j$
2. d-choices GC algorithm selects $d \geq 2$ blocks uniformly at random and erases a block containing the smallest number of valid pages among the d selected blocks:

$$p_j(\vec{m}) = \left(\sum_{\ell=j}^{b} m_\ell\right)^d - \left(\sum_{\ell=j+1}^{b} m_\ell\right)^d$$

3. Greedy GC algorithm: d-choices with $d = N$.

Universiteit Antwerpen

Workload model

Rosenblum model

- A fraction f of the data is termed hot
- Hot pages are updated at rate $r \geq f$, cold pages at rate $1 - r$
- Reducing f or increasing r makes hot data hotter

Special case: Uniform random writes ($r = f$)

Every (logical) page on the device is updated with the same probability

Universiteit Antwerpen

System description

GC mechanism: Double Write Frontier (DWF)

- Uses 2 write frontiers:
  - WFE: External WF for externally issued writes (by host)
  - WFI: Internal WF used during GC
- Separates data without hot/cold data identification techniques

Garbage collection with DWF

- GC algorithm invoked when WFE becomes full, chooses victim containing j valid pages
- Assume WFI contains $b - j^*$ pages in erase state
  - $j \leq b - j^*$: j valid pages copied to WFI, victim becomes new WFE
  - $j > b - j^*$: j valid pages copied to WFI, victim becomes new WFI, reinvoke GC

Universiteit Antwerpen

Performance measures

PE fairness

- Mean number of PE cycles performed on blocks before any block reaches W_{max} PE cycles, divided by W_{max}
- Describes how fairly the PE cycles are distributed among blocks

SSD endurance

- Expected number of external writes before any block reaches W_{max} PE cycles
- Expresses the life span of the device in full drive writes (FDW)

Main questions

- How much can wear leveling mechanisms improve performance?
- Which of the proposed measures dominates w.r.t. performance and what role does the GCA play?

Model framework

Background on mean field models

- Stochastic system of N interacting blocks (N-dimensional Markov chain)
- Problem: impractical to compute steady state for large N
- Solution: consider the limit of N tending to infinity
- Limit is a deterministic system, its evolution captured by the trajectories of a set of ODEs (called drift equations)
- Drift corresponds to studying the behavior of one (type of) block, averaging the effects of other blocks

Universiteit Antwerpen

Model framework

Drift equations and fixed point (for uniform random writes)

- Let $f_{i,w}(\vec{m})$ represent the expected change in the fraction of blocks containing i valid pages with w PE cycles
- Determine fixed point $\vec{m}^\star$ where

$$\sum_{i=0}^{b} f_{i,w}(\vec{m}^\star) = 0$$

- WA, PE fairness and SSD endurance based on fixed point
- Gives exact results for N tending to infinity (provided that limits are exchangeable)

Model extension for Rosenblum workload

- Can be extended for hot/cold workload, but numerical solution is computation intensive

Main findings: Uniform random writes

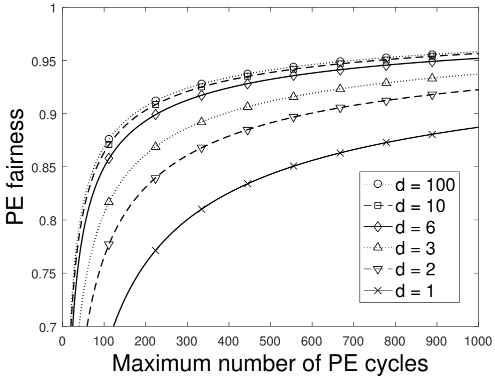

Figure: PE fairness under uniform random writes with $b = 32$ and $S_f = 0.1$.

Increasing b, d or S_f results in increased PE fairness.
Narrow margin for improving by implementing wear leveling mechanisms.

Universiteit Antwerpen

Main findings: Uniform random writes

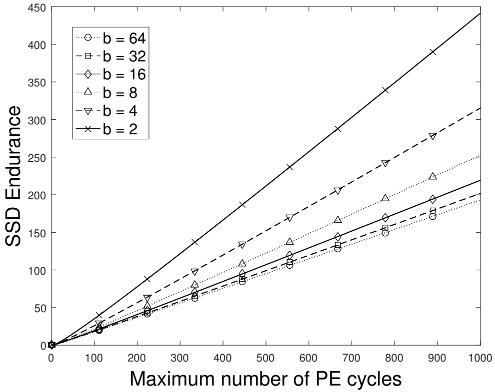

Figure: SSD endurance under uniform random writes with $S_f = 0.1$ and $d = 10$.

Smaller block sizes b result in higher endurance due to lower WA.

Universiteit Antwerpen

Main findings: Hot/cold data

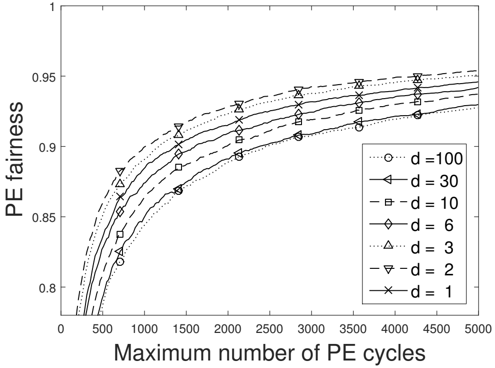

Figure: PE fairness under hot/cold data (DWF) with $b = 32$, $S_f = 0.1$, $f = 0.2$ and $r = 0.8$.

In contrast with uniform random writes, smaller d and S_f values result in better fairness.
The GCA then more likely chooses victims containing primarily cold data, which have a lower number of PE cycles.

Main findings: Hot/cold data

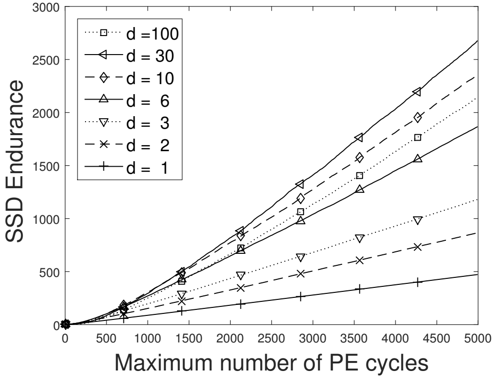

Figure: SSD endurance under hot/cold data (DWF) with $b = 32$, $S_f = 0.1$, $f = 0.2$ and $r = 0.8$.

There exists a finite d value optimizing SSD endurance. Smaller d have higher WA, outweighing benefits of better PE fairness.

Universiteit Antwerpen

Main findings: Hot/cold data

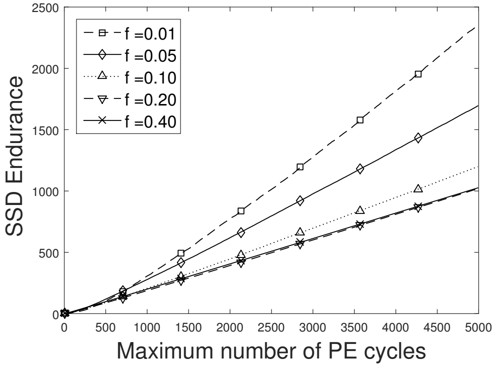

Figure: SSD endurance under hot/cold data (DWF) with $b = 32$, $S_f = 0.1$ and $d = 10$.

PE fairness reduces with hotness, but lower WA of hotter data results in better endurance.

Main findings: Hot/cold data

ρ	r	f	$d_{\text{MIN WA}}$	$d_{\text{MAX End}}$	$d_{\text{MAX Fair}}$
0.90	0.80	0.20	13	13	2
0.90	0.95	0.05	10	8	1
0.90	0.99	0.01	27	24	1
0.85	0.80	0.20	11	9	2
0.90	0.80	0.20	13	13	2
0.95	0.80	0.20	22	21	2

Table: Comparison of d values optimizing WA, SSD endurance and PE fairness for several parameter settings in a system of $N = 10,000$ blocks of size $b = 32$ with $W_{max} = 5000$ (10 runs).

Minimizing WA is more beneficial for endurance than maximizing fairness.

Universiteit Antwerpen

Main takeaway

Uniform random writes

- Greedy (or *d* large) GC delivers near optimal fairness and endurance
- Large blocksizes can result in shorter life span (high WA outweighs fairness)

Hot/cold data (DWF)

- Increasing data hotness leads to lower PE fairness
- Lowering PE fairness may lead to higher SSD endurance
- *d* values maximizing endurance are relatively small, but closer to those minimizing WA than those maximizing PE fairness
- When increasing hotness, minimizing WA outweighs achieving roughly equal wear

Universiteit Antwerpen

Possible extensions and ongoing work

Possible extensions

- GC algorithms depending on wear of blocks
- Impact of wear leveling mechanism on SSD endurance

Ongoing and future work

- Fairness and endurance of trace-based workloads
- Impact of data separation techniques on device lifespan

Universiteit Antwerpen